Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors

Scott E. Carrell

University of California, Davis and National Bureau of Economic Research

James E. West

U.S. Air Force Academy

In primary and secondary education, measures of teacher quality are often based on contemporaneous student performance on standardized achievement tests. In the postsecondary environment, scores on student evaluations of professors are typically used to measure teaching quality. We possess unique data that allow us to measure relative student performance in mandatory follow-on classes. We compare metrics that capture these three different notions of instructional quality and present evidence that professors who excel at promoting contemporaneous student achievement teach in ways that improve their student evaluations but harm the follow-on achievement of their students in more advanced classes.

Thanks go to U.S. Air Force Academy personnel, R. Schreiner, W. Bremer, R. Fullerton, J. Putnam, D. Stockburger, K. Carson, and P. Egleston, for assistance in obtaining the data for this project and to Deb West for many hours entering data from archives. Thanks also go to F. Hoffmann, H. Hoynes, C. Hoxby, S. Imberman, C. Knittel, L. Lefgren, M. Lovenheim, T. Maghakian, D. Miller, P. Oreopoulos, M. Page, J. Rockoff, and D. Staiger and all seminar participants at the American Education Finance Association meetings, Clemson University, Duke University, NBER Higher Ed Working Group, Stanford University, and University of California, Berkeley and Davis for their helpful comments. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the U.S. Air Force, the Department of Defense, or the U.S. government.

[*Journal of Political Economy*, 2010, vol. 118, no. 3]

> A weak faculty operates a weak program that attracts weak students. (Koerner 1963)

I. Introduction

Conventional wisdom holds that "higher-quality" teachers promote better educational outcomes. Since teacher quality cannot be directly observed, measures have largely been driven by data availability. At the elementary and secondary levels, scores on standardized student achievement tests are the primary measure used and have been linked to teacher bonuses and terminations (Figlio and Kenny 2007). At the postsecondary level, student evaluations of professors are widely used in faculty promotion and tenure decisions. However, teachers can influence these measures in ways that may reduce actual student learning. Teachers can "teach to the test." Professors can inflate grades or reduce academic content to elevate student evaluations. Given this, how well do each of these measures correlate with the desired outcome of actual student learning?

Studies have found mixed evidence regarding the relationship between observable teacher characteristics and student achievement at the elementary and secondary education levels.[1] As an alternative method, teacher "value-added" models have been used to measure the total teacher input (observed and unobserved) to student achievement. Several studies find that a one-standard-deviation increase in teacher quality improves student test scores by roughly one-tenth of a standard deviation (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Aaronson et al. 2007; Kane, Rockoff, and Staiger 2008). However, recent evidence from Kane and Staiger (2008) and Jacob, Lefgren, and Sims (2010) suggests that these contemporaneous teacher effects may decay relatively quickly over time,[2] and Rothstein (2010) finds evidence that the nonrandom place-

[1] Jacob and Lefgren (2004) find that principal evaluations of teachers were the best predictor of student achievement; Clotfelter, Ladd, and Vigdor (2006, 2007) find evidence that National Board certification and teacher licensure test scores positively predict teacher effectiveness; Dee (2004, 2005) finds that students perform better with same race and gender teachers; and Harris and Sass (2007) find some evidence that teacher professional development is positively correlated with student achievement in middle and high school math. Summers and Wolfe (1977), Cavalluzzo (2004), Vandevoort, Amrein-Beardsley, and Berliner (2004), and Goldhaber and Anthony (2007) find positive effects from teachers certified by the National Board for Professional Teaching Standards. See also Hanushek (1971), Murnane (1975), Summers and Wolfe (1977), Ehrenberg and Brewer (1994), Ferguson and Ladd (1996), Boyd et al. (2006), and Aaronson, Barrow, and Sander (2007).

[2] Jacob et al. (2010) find that 20 percent of the contemporaneous effects persist into the subsequent year. Rothstein (2010) finds that roughly 50 percent persists into year 1 and none persists into year 2 for mathematics courses.

ment of students to teachers may bias value-added estimates of teacher quality.[3]

Even less is known about how the quality of instruction affects student outcomes at the postsecondary level.[4] Standardized achievement tests are not given at the postsecondary level, and grades are not typically a consistent measure of student academic achievement because of heterogeneity of assignments/exams and the mapping of those assessment tools into final grades across individual professors. Additionally, it is difficult to measure how professors affect student achievement because students generally "self-select" their course work and their professors. For example, if better students tend to select better professors, then it is difficult to statistically separate the teacher effects from the selection effects. As a result, the primary tool used by administrators to measure professor teaching quality is scores on subjective student evaluations, which are likely endogenous with respect to (expected) student grades.

To address these various measurement and selection issues in measuring teacher quality, our study uses a unique panel data set from the United States Air Force Academy (USAFA) in which students are randomly assigned to professors over a wide variety of standardized core courses. The random assignment of students to professors, along with a vast amount of data on both professors and students, allows us to examine how professor quality affects student achievement free from the usual problems of self-selection. Furthermore, performance in USAFA core courses is a consistent measure of student achievement because faculty members teaching the same course use an identical syllabus and give the same exams during a common testing period.[5] Finally, USAFA students are required to take and are randomly assigned to numerous follow-on courses in mathematics, humanities, basic sciences, and engineering. Performance in these mandatory follow-on courses is arguably a more persistent measurement of student learning. Thus, a distinct advantage of our data is that even if a student has a particularly poor introductory course professor, he or she still is required to take the follow-on related curriculum.[6]

[3] However, Kane and Staiger (2008) show that controlling for prior year test scores produces unbiased estimates in the presence of self-selection.

[4] Hoffmann and Oreopoulos (2009) find that perceived professor quality, as measured by teaching evaluations, affects the likelihood of a student dropping a course and taking subsequent courses in the same subject. Other recent postsecondary studies have focused on the effectiveness of part-time (adjunct) professors. See Ehrenberg and Zhang (2005) and Bettinger and Long (2006).

[5] Common testing periods are used for freshman- and sophomore-level core courses. All courses are taught without the use of teaching assistants, and faculty members are required to be available for appointments with students from 7:30 a.m. to 4:30 p.m. each day classes are in session.

[6] For example, students of particularly bad Calculus I instructors must still take Calculus II and six engineering courses, even if they decide to be a humanities major.

These properties enable us to measure professor quality free from selection and attrition bias. We start by estimating professor quality using teacher value-added in the contemporaneous course. We then estimate value-added for subsequent classes that require the introductory course as a prerequisite and examine how these two measures covary. That is, we estimate whether high- (low-) value-added professors in the introductory course are high- (low-) value-added professors for student achievement in follow-on related curriculum. Finally, we examine how these two measures of professor value-added (contemporaneous and follow-on achievement) correlate with professor observable attributes and student evaluations of professors. These analyses give us a unique opportunity to compare the relationship between value-added models (currently used to measure primary and secondary teacher quality) and student evaluations (currently used to measure postsecondary teacher quality).

Results show that there are statistically significant and sizable differences in student achievement across introductory course professors in both contemporaneous and follow-on course achievement. However, our results indicate that professors who excel at promoting contemporaneous student achievement, on average, harm the subsequent performance of their students in more advanced classes. Academic rank, teaching experience, and terminal degree status of professors are negatively correlated with contemporaneous value-added but positively correlated with follow-on course value-added. Hence, students of less experienced instructors who do not possess a doctorate perform significantly better in the contemporaneous course but perform worse in the follow-on related curriculum.

Student evaluations are positively correlated with contemporaneous professor value-added and negatively correlated with follow-on student achievement. That is, students appear to reward higher grades in the introductory course but punish professors who increase deep learning (introductory course professor value-added in follow-on courses). Since many U.S. colleges and universities use student evaluations as a measurement of teaching quality for academic promotion and tenure decisions, this latter finding draws into question the value and accuracy of this practice.

These findings have broad implications for how students should be assessed and teacher quality measured. Similar to elementary and secondary school teachers, who often have advance knowledge of assessment content in high-stakes testing systems, all professors teaching a given course at USAFA have an advance copy of the exam before it is given. Hence, educators in both settings must choose how much time to allocate to tasks that have great value for raising current scores but may have little value for lasting knowledge. Using our various measures

of quality to rank-order professors leads to profoundly different results. As an illustration, the introductory calculus professor in our sample who ranks dead last in deep learning ranks sixth and seventh best in student evaluations and contemporaneous value-added, respectively. These findings support recent research by Barlevy and Neal (2009), who propose an incentive pay scheme that links teacher compensation to the ranks of their students within appropriately defined comparison sets and requires that new assessments consisting of entirely new questions be given at each testing date. The use of new questions eliminates incentives for teachers to coach students concerning the answers to specific questions on previous assessments.

The remainder of the paper proceeds as follows: Section II reviews the empirical setting. Section III presents the methods and results for professor value-added models. Section IV examines how the observable attributes of professors and student evaluations of instructors are correlated with professor value-added. Section V presents concluding remarks.

II. Empirical Setting

The U.S. Air Force Academy is a fully accredited undergraduate institution of higher education with an approximate enrollment of 4,500 students. There are 32 majors offered including the humanities, social sciences, basic sciences, and engineering. Applicants are selected for admission on the basis of academic, athletic, and leadership potential. All students attending USAFA receive a 100 percent scholarship to cover their tuition, room, and board. Additionally, each student receives a monthly stipend of $845 to cover books, uniforms, computer, and other living expenses. All students are required to graduate within 4 years[7] and serve a 5-year commitment as a commissioned officer in the U.S. Air Force following graduation.

Approximately 40 percent of classroom instructors at USAFA have terminal degrees, as one might find at a university where introductory course work is often taught by graduate student teaching assistants. However, class sizes are very small (average of 20), and student interaction with faculty members is encouraged. In this respect, students' learning experiences at USAFA more closely resemble those of students who attend small liberal arts colleges.

Students at USAFA are high achievers, with average math and verbal Scholastic Aptitude Test (SAT) scores at the 88th and 85th percentiles

[7] Special exceptions are given for religious missions, medical "setbacks," and other instances beyond the control of the individual.

of the nationwide SAT distribution.[8] Students are drawn from each congressional district in the United States by a highly competitive process, ensuring geographic diversity. According to the National Center for Education Statistics (http://nces.ed.gov/globallocator/), 14 percent of applicants were admitted to USAFA in 2007. Approximately 17 percent of the sample is female, 5 percent is black, 7 percent is Hispanic, and 6 percent is Asian. Twenty-six percent are recruited athletes, and 20 percent attended a military preparatory school. Seven percent of students at USAFA have a parent who graduated from a service academy and 17 percent have a parent who previously served in the military.

A. The Data Set

Our data set consists of 10,534 students who attended USAFA from the fall of 2000 through the spring of 2007. Student-level pre-USAFA data include whether students were recruited as athletes, whether they attended a military preparatory school, and measures of their academic, athletic, and leadership aptitude. Academic aptitude is measured through SAT verbal and SAT math scores and an academic composite computed by the USAFA admissions office, which is a weighted average of an individual's high school grade point average (GPA), class rank, and the quality of the high school attended. The measure of pre-USAFA athletic aptitude is a score on a fitness test required by all applicants prior to entrance.[9] The measure of pre-USAFA leadership aptitude is a leadership composite computed by the USAFA admissions office, which is a weighted average of high school and community activities (e.g., student council officer, Eagle Scout, captain of a sports team, etc.).

Our primary outcome measure consists of a student-level census of all courses taken and the percentage of points earned in each course. We normalize the percentage of points earned within a course/semester to have a mean of zero and a standard deviation of one. The average percentage of points earned in the course is 78.17, which corresponds to a mean GPA of 2.75.

Students at USAFA are required to take a core set of approximately 30 courses in mathematics, basic sciences, social sciences, humanities, and engineering.[10] Table 1 provides a list of the required math, science, and engineering core courses.

[8] See http://professionals.collegeboard.com/profdownload/sat_percentile_ranks_2008.pdf for SAT score distributions.

[9] Barron, Ewing, and Waddell (2000) found a positive correlation between athletic participation and educational attainment, and Carrell, Fullerton, and West (2009) found a positive correlation between fitness scores and academic achievement.

[10] Over the period of our study there were some changes made to the core curriculum at USAFA.

TABLE 1
REQUIRED MATH AND SCIENCE CORE CURRICULUM

Course	Description	Credit Hours
Basic sciences:		
Biology 215	Introductory Biology with Lab	3
Chemistry 141 and 142 or 222	Applications of Chemistry I and II	6
Computer Science 110	Introduction to Computing	3
Mathematics 141	Calculus I	3
Mathematics 142 or 152[a]	Calculus II	3
Mathematics 300, 356, or 377[a]	Introduction to Statistics	3
Physics 110[a]	General Physics I	3
Physics 215[a]	General Physics II	3
Engineering:		
Engineering 100	Introduction to Engineering Systems	3
Engineering 210[a]	Civil Engineering—Air Base Design and Performance	3
Engineering Mechanics 120[a]	Fundamentals of Mechanics	3
Aeronautics 315[a]	Fundamentals of Aeronautics	3
Astronautics 310[a]	Introduction to Astronautics	3
Electrical Engineering 215 or 231[a]	Electrical Signals and Systems	3
Total		45

[a] Denotes that Calculus I is required as a prerequisite to the course.

Individual professor-level data were obtained from USAFA historical archives and the USAFA Center for Education Excellence and were matched to the student achievement data for each course taught by section-semester-year.[11] Professor data include academic rank, gender, education level (master of arts or doctorate), years of teaching experience at USAFA, and scores on subjective student evaluations.

Over the 10-year period of our study we estimate our models using student performance across 2,820 separate course-sections taught by 421 different faculty members. Average class size was 20 students, and approximately 38 sections of each course were taught per year. The average number of classes taught by each professor in our sample is nearly seven. Table 2 provides summary statistics of the data.

[11] Owing to the sensitivity of the data, we were able to obtain the professor observable data only for mathematics, chemistry, and physics. Results for physics and chemistry professors can be found in Carrell and West (2008). Because of the large number of faculty in these departments, a set of demographic characteristics (e.g., female assistant professor, doctorate with 3 years of experience) does not uniquely identify an individual faculty member.

TABLE 2
SUMMARY STATISTICS

	Observations	Mean	Standard Deviation
Student-level variables:			
Total course hours	10,534	16.29	7.99
GPA	10,534	2.75	.80
Percentage of points earned in courses (mean)	10,534	78.17	8.45
SAT verbal	10,534	632.30	66.27
SAT math	10,534	663.51	62.80
Academic composite	10,533	12.82	2.13
Leadership composite	10,508	17.30	1.85
Fitness score	10,526	4.66	.99
Female	10,534	.17	.38
Black	10,534	.05	.22
Hispanic	10,534	.07	.25
Asian	10,534	.06	.23
Recruited athlete	10,534	.25	.44
Attended preparatory school	10,534	.20	.40
Professor-level variables:[a]			
Instructor is a lecturer	91	.58	.50
Instructor is an assistant professor	91	.26	.44
Instructor is an associate or full professor	91	.15	.36
Instructor has a terminal degree	91	.31	.46
Instructor's teaching experience	91	3.66	4.42
Number of sections taught	421	6.64	5.22
Class-level variables:[b]			
Class size	2,820	20.28	3.48
Number of sections per course per year	2,820	38.37	12.80
Average class SAT verbal	2,820	631.83	22.05
Average class SAT math	2,820	661.61	27.77
Average class academic composite	2,820	12.84	.73
Student evaluation of professors by section:[c]			
Instructor's ability to provide clear, well-organized instruction was	237	4.48	.70
Value of questions and problems raised by instructor was	237	4.50	.57
Instructor's knowledge of course material was	237	5.02	.58
The course as a whole was	237	4.08	.61
Amount you learned in the course was	237	4.09	.58
The instructor's effectiveness in facilitating my learning in the course was	237	4.42	.69

[a] Observable attribute data are available only for calculus professors.
[b] Class-level data include introductory calculus and follow-on related core courses.
[c] Student evaluation data are for introductory calculus professors only. The number of observations is the number of sections.

B. Student Placement into Courses and Sections

Prior to the start of the freshman academic year, students take course placement exams in mathematics, chemistry, and select foreign languages. Scores on these exams are used to place students into the appropriate starting core courses (i.e., remedial math, Calculus I, Calculus II, etc.). Conditional on course placement, the USAFA registrar employs a stratified random assignment algorithm to place students into sections within each course/semester. The algorithm first assigns all female students evenly throughout all offered sections, then places male-recruited athletes, and then assigns all remaining students. Within each group (i.e., female, male athlete, and all remaining males), assignments are random with respect to academic ability and professor.[12] Thus, students throughout their 4 years of study have no ability to choose their professors in required core courses. Faculty members teaching the same course use an identical syllabus and give the same exams during a common testing period. These institutional characteristics assure that there is no self-selection of students into (or out of) courses or toward certain professors.

Although the placement algorithm used by the USAFA registrar should create sections that are a random sample of the course population with respect to academic ability, we employed resampling techniques as in Lehmann and Romano (2005) and Good (2006) to empirically test this assumption. For each section of each core course/semester we randomly drew 10,000 sections of equal size from the relevant introductory course enrollment without replacement. Using these randomly sampled sections, we computed the sums of both the academic composite score and the SAT math score.[13] We then computed empirical p-values for each section, representing the proportion of simulated sections with values less than that of the observed section.

Under random assignment, any unique p-value is equally likely to be observed; hence the expected distribution of the empirical p-values is uniform. We tested the uniformity of the distributions of empirical p-values by semester by course using both a Kolmogorov-Smirnov one-sample equality of distribution test and a χ^2 goodness of fit test.[14] As

[12] In-season intercollegiate athletes are not placed into the late-afternoon section, which starts after 3:00 p.m.

[13] We performed resampling analysis on the USAFA classes of 2000–2009. We also conducted the resampling analysis for SAT verbal and math placement scores and found qualitatively similar results. For brevity we do not present these results in the text.

[14] The Kolmogorov-Smirnov test equals $\sup_x |F_n(x) - F(x)|$, where $F_n(x)$ is the empirical cumulative distribution function and $F(x)$ is the theoretical cumulative distribution function;

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \eta_i)^2}{\eta_i},$$

where n_i is the observed frequency in bin i and η_i is the expected frequency in bin i.

TABLE 3
RANDOMNESS CHECKS

	CALCULUS I		CALCULUS II	
PROFESSOR CHARACTERISTIC	Academic Composite (1)	SAT Math (2)	Academic Composite (1)	SAT Math (2)
Associate/full professor	.029	.002	.009	−.024
	(.060)	(.073)	(.060)	(.056)
Experience	.000	.000	.002	−.003
	(.005)	(.006)	(.007)	(.004)
Terminal degree	.031	.056	.038	−.006
	(.039)	(.040)	(.042)	(.037)
Empirical *p*-values (mean and standard deviation)	.512	.514	.503	.503
	(.311)	(.334)	(.302)	(.315)
Kolmogorov-Smirnov test (no. failed/total tests)	0/20	1/20	0/20	0/20
χ^2 goodness of fit test (no. failed/total tests)	1/20	2/20	0/20	0/20

NOTE.—Each cell represents regression results in which the dependent variable is the empirical *p*-value from resampling as described in Sec. II.B and the independent variable is the professor characteristic. Because of the collinearity of the regressors, each column represents results for three separate regressions for associate/full professor, experience, and terminal degree. All specifications include semester by year fixed effects. Standard errors are clustered by professor. The empirical *p*-value of each section represents the proportion of the 10,000 simulated sections with values less than that of the observed section. The Kolmogorov-Smirnov and χ^2 goodness of fit test results indicate the number of tests of the uniformity of the distribution of *p*-values that failed at the 5 percent level.

reported in table 3, we rejected the null hypothesis of random placement for only one of 80 course/semester test statistics at the .05 level using the Kolmogorov-Smirnov test and three of 80 course/semester test statistics using the χ^2 goodness of fit test. As such, we found virtually no evidence of nonrandom placement of students into sections by academic ability.

Next, we tested for the random placement of professors with respect to student ability by regressing the empirical *p*-values from resampling by section on professor academic rank, years of experience, and terminal degree status. Results for this analysis are shown in table 3 and indicate that there is virtually no evidence of nonrandom placement of professors into course sections. Of the 36 estimated coefficients, none are statistically significant at the 5 percent level.

Results from the preceding analyses indicate that the algorithm that places students into sections within a course and semester appears to be random with respect to both student and professor characteristics.

C. Are Student Scores a Consistent Measure of Student Achievement?

The integrity of our results depends on the percentage of points earned in core courses being a consistent measure of relative achievement across students. The manner in which student scores are determined at USAFA,

particularly in the Math Department, allows us to rule out potential mechanisms for our results. Math professors grade only a small proportion of their own students' exams, vastly reducing the ability of "easy" or "hard" grading professors to affect their students' scores. All math exams are jointly graded by all professors teaching the course during that semester in "grading parties," where Professor A grades question 1 and Professor B grades question 2 for all students taking the course. These aspects of grading allow us to rule out the possibility that professors have varying grading standards for equal student performance. Hence, our results are likely driven by the manner in which the course is taught by each professor.

In some core courses at USAFA, 5–10 percent of the overall course grade is earned by professor/section-specific quizzes and/or class participation. However, for the period of our study, the introductory calculus course at USAFA did not allow for any professor-specific assignments or quizzes. Thus, potential "bleeding heart" professors had no discretion to boost grades or to keep their students from failing their courses. For this reason, we present results in this study for the introductory calculus course and follow-on courses that require introductory calculus as a prerequisite.[15]

III. Professor Value-Added

A. *Empirical Model*

The professor value-added model estimates the total variance in professor inputs (observed and unobserved) in student academic achievement by utilizing the panel structure of our data, where different professors teach multiple sections of the same course across years. We estimate professor value-added using a random effects model. Random effects estimators are minimum variance and efficient but are not typically used in the teacher quality literature because of the stringent requirement for consistency—that teacher value-added be uncorrelated with all other explanatory variables in the model.[16] This requirement is almost certainly violated when students self-select into course work or sections of a given course, but the requirement is satisfied in our context (Raudenbush and Bryk 2002; McCaffrey et al. 2004).

Consider a set of students indexed by $i = 1, \ldots, N$ who are randomly placed into sections $s^1 \in \mathbb{S}$ of the introductory course, where the su-

[15] We find qualitatively similar results for chemistry and physics professors in Carrell and West (2008), where the identification is less clean. Chemistry and physics professors were allowed to have section-specific assignments and grade their own students' exams. These results are available on request.

[16] We run a Hausman specification test and fail to reject the null hypothesis that the fixed effects and random effects estimates are equivalent.

perscript 1 denotes an introductory course section. A member of the set of introductory course professors, indexed by $j^1 = 1, \ldots, J$, is assigned to each section s^1. In subsequent semesters, each student i is randomly placed into follow-on course sections $s^2 \in \mathbb{S}$, where the superscript 2 denotes a follow-on course section. A member of the set of follow-on course professors, indexed by $j^2 = 1, \ldots, J$ (overlapping the set of introductory course professors), is assigned to each section s^2.

The outcomes of student i are given by the following two-equation model:

$$\begin{bmatrix} Y^1_{itj^1j^2s^1s^2} \\ Y^2_{itj^1j^2s^1s^2} \end{bmatrix} = \begin{bmatrix} X_{its^1} & 0 \\ 0 & X_{its^2} \end{bmatrix} \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix} + \begin{bmatrix} \gamma_t^1 \\ \gamma_t^2 \end{bmatrix} + \begin{bmatrix} \lambda^1_{j^1} + \lambda^1_{j^2} + \xi^1_{ts^1} + \xi^1_{ts^2} \\ \lambda^2_{j^2} + \lambda^2_{j^1} + \xi^2_{ts^2} + \xi^2_{ts^1} \end{bmatrix} + \begin{bmatrix} \epsilon^1_{itj^1j^2s^1s^2} \\ \epsilon^2_{itj^1j^2s^1s^2} \end{bmatrix}, \quad (1)$$

where $Y^1_{itj^1j^2s^1s^2}$ and $Y^2_{itj^1j^2s^1s^2}$ are the normalized percentages of points earned by student i in semester-year t with introductory professor j^1 in section s^1 and follow-on professor j^2 in section s^2. Superscript 1 denotes introductory course achievement and superscript 2 denotes follow-on course achievement. The terms X_{its^1} and X_{its^2} are vectors of student-specific and classroom mean peer characteristics, including SAT math, SAT verbal, academic composite, fitness score, leadership composite, race/ethnicity, gender, recruited athlete, and whether they attended a military preparatory school relevant to sections s^1 and s^2, respectively, in time t. We control for unobserved mean differences in academic achievement or grading standards across time by including course by semester intercepts, γ_t^1 and γ_t^2.

The λ's are the parameters of primary interest in our study, which measure professor value-added. Specifically, $\lambda^1_{j^1}$ measures the introductory course professor j^1's value-added in the contemporaneous introductory course and $\lambda^2_{j^1}$ measures the introductory course professor j^1's value-added in mandatory follow-on related courses (deep learning). Likewise, $\lambda^2_{j^2}$ measures the follow-on course professor j^2's value-added in the contemporaneous follow-on course and $\lambda^1_{j^2}$ measures the follow-on course professor j^2's value-added in the introductory course. The presence of $\lambda^1_{j^2}$ allows for a second test of random assignment since we expect this effect to be zero. High values of λ indicate that the professor's students perform better on average, and low values of λ indicate lower average achievement. The variance of λ across professors measures the dispersion of professor quality, whether it be observed or unobserved (Rivkin et al. 2005).

The ξ terms are section-specific random effects measuring classroom-level common shocks that are independent across professors j and time t. Specifically, $\xi^1_{ts^1}$ measures the introductory course section-specific

shock in the contemporaneous introductory course, and $\xi^{2}_{ts^{1}}$ measures the introductory course section-specific common shock in the follow-on course. Likewise, $\xi^{2}_{ts^{2}}$ measures the follow-on course section-specific shock in the contemporaneous follow-on course, and $\xi^{1}_{ts^{2}}$ measures the follow-on course section-specific common shock in the introductory course. Again, we expect this latter effect to be zero given the random assignment of students to follow-on course sections.

The terms $\epsilon^{1}_{itj^{1}j^{2}s^{1}s^{2}}$ and $\epsilon^{2}_{itj^{1}j^{2}s^{1}s^{2}}$ are the student-specific stochastic error terms in the introductory and follow-on course, respectively.[17]

B. *Results for Introductory Professors*

Table 4 presents the full set of estimates of the variances and covariances of the λ's, ξ's, and ϵ's for introductory calculus professors. Covariance elements in the matrix with a value of 0 were set to zero in the model specification.[18]

The estimated variance in introductory professor quality in the contemporaneous introductory course, Var $(\lambda^{1}_{j^{1}})$ in row 1, column 1, is 0.0028 (standard deviation [SD] = 0.052) and is statistically significant at the .05 level. This result indicates that a one-standard-deviation change in professor quality results in a 0.05-standard-deviation change in student achievement. In terms of scores, this effect translates into about 0.6 percent of the final percentage of points earned in the course. The magnitude of the effect is slightly smaller but qualitatively similar to those found in elementary school teacher quality estimates (Kane et al. 2008).

When evaluating achievement in the contemporaneous course being taught, the major threat to identification is that the professor value-added model could be identifying a common treatment effect rather than measuring the true quality of instruction. For example, if Professor A "teaches to the test," his students may perform better on exams and earn higher grades in the course, but they may not have learned any more actual knowledge relative to Professor B, who does not teach to the test. In the aforementioned scenario, the contemporaneous model would identify Professor A as a higher-quality teacher than Professor B.

[17] Owing to the complexity of the nesting structure of professors within courses and course sections within professors, we estimate all the above parameters in two separate random effects regression models using Stata's xtmixed command—one model for introductory course professors and another for follow-on course professors.

[18] A unique aspect of our data is that we observe the same professors teaching multiple sections of the same course in each year. In results unreported but available on request, we tested the stability of professor value-added across years and found insignificant variation in the within-professor teacher value-added across years. These results indicate that the existing practice in the teacher quality literature of relying on only year-to-year variation appears to be justified in our setting.

TABLE 4

VARIANCE-COVARIANCE OF PROFESSOR VALUE-ADDED AND COURSE SECTIONS IN CONTEMPORANEOUS AND FOLLOW-ON COURSES

	λ^1_{j1} (1)	λ^2_{j1} (2)	λ^1_{j2} (3)	λ^2_{j2} (4)	ξ^1_{ts1} (5)	ξ^2_{ts1} (6)	ξ^1_{ts2} (7)	ξ^2_{ts2} (8)	ϵ (9)
1. λ^1_{j1}	.0028								
	(.0001, .0538)								
2. λ^2_{j1}	−.0004	.0025							
	(−.0051, .0043)	(.0006, .0115)							
3. λ^1_{j2}	0	0	.0000						
			(.0000, .0000)						
4. λ^2_{j2}	0	0	0	.0186					
				(.0141, .0245)					
5. ξ^1_{ts1}	0	0	0	0	.0255				
					(.0159, .0408)				
6. ξ^2_{ts1}	0	0	0	0	.0251	.0248			
					(.0179, .0324)	(.0196, .0315)			
7. ξ^1_{ts2}	0	0	0	0	0	0	.0000		
							(.0000, .0000)		
8. ξ^2_{ts2}	0	0	0	0	0	0	0	.0100	
								(.0067, .0149)	
9. ϵ	0	0	0	0	0	0	0	0	.7012
									(.6908, .7117)

NOTE.—The table shows random effects estimates of the variances and covariances for professors, course sections, and students. The model specification includes course by semester fixed effects as well as classroom-level attributes for SAT math, SAT verbal, and academic composite. Individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score and their interactions with a follow-on course indicator. Ninety-five percent confidence intervals are shown in parentheses. The term λ is the random effect of the intro (subscript j^1) or follow-on (subscript j^2) professor in the intro (superscript 1) or follow-on (superscript 2) course; ξ is the random effect of the intro (subscript s^1) or follow-on (subscript s^2) section in the intro (superscript 1) or follow-on (superscript 2) course; and ϵ is the course achievement level error term.

The USAFA's comprehensive core curriculum provides a unique opportunity to test how introductory course professors affect follow-on course achievement free from selection bias. The estimate of Var $(\lambda^2_{j^1})$ is shown in row 2, column 2 of table 4 and indicates that introductory course professors significantly affect follow-on course achievement.[19] The variance in follow-on course value-added is estimated to be 0.0025 (SD = 0.050). The magnitude of this effect is roughly equivalent to that estimated in the contemporaneous course and indicates that a one-standard-deviation change in introductory professor quality results in a 0.05-standard-deviation change in follow-on course achievement.

The preceding estimates of Var $(\lambda^1_{j^1})$ and of Var $(\lambda^2_{j^1})$ indicate that introductory course calculus professors significantly affect student achievement in both the contemporaneous introductory course being taught and follow-on courses. The estimated covariance, Cov $(\lambda^1_{j^1}, \lambda^2_{j^1})$, of these professor effects is negative (-0.0004) and statistically insignificant as shown in column 1, row 2 of table 4. This result indicates that being a high- (low-) value-added professor for contemporaneous student achievement is negatively correlated with being a high- (low-) value-added professor for follow-on course achievement. To get a better understanding of this striking result, we next decompose the covariance estimate.

We note that there are two ways in which the introductory professor (i.e., introductory calculus professor) can affect follow-on course achievement (i.e., aeronautical engineering). First, the initial course professor effect can persist into the follow-on course, which we will specify as $\rho\lambda^1_{j^1}$. Second, the initial course professor can produce value-added not reflected in the initial course, which we will specify as $\phi^2_{j^1}$. One example of $\phi^2_{j^1}$ would be "deep learning" or understanding of mathematical concepts that are not measured on the calculus exam but would increase achievement in more advanced mathematics and engineering courses. Hence, we can specify $\lambda^2_{j^1}$ and its estimated covariance with $\lambda^1_{j^1}$ as follows:[20]

$$\lambda^2_{j^1} = \rho\lambda^1_{j^1} + \phi^2_{j^1}, \tag{2}$$

$$\begin{aligned}\mathbb{E}[\lambda^1_{j^1}\lambda^2_{j^1}] &= \mathbb{E}[(\lambda^1_{j^1})(\rho\lambda^1_{j^1} + \phi^2_{j^1})] \\ &= \rho\,\mathrm{Var}\,(\lambda^1_{j^1}).\end{aligned} \tag{3}$$

Therefore, Cov $(\lambda^1_{j^1}, \lambda^2_{j^1})/$ Var $(\lambda^1_{j^1})$ is a consistent estimate of ρ, the pro-

[19] We estimate $\lambda^2_{j^1}$ using all the follow-on required courses that require Calculus I as a prerequisite. These courses are listed in table 1.

[20] If $\phi^2_{j^1}$ represents value-added from the initial course professor in the follow-on course not reflected in initial course achievement, Cov $(\lambda^1_{j^1}, \phi^2_{j^1}) = 0$ by construction.

portion of contemporaneous value-added that persists into follow-on course achievement.

Using results from table 4, we estimate ρ at -0.14.[21] Taken jointly, our estimates of Var $(\lambda_{j^1}^1)$, Var $(\lambda_{j^1}^2)$, and ρ indicate that one set of calculus professors produce students who perform relatively better in calculus and another set of calculus professors produce students who perform well in follow-on related courses, and these sets of professors are not the same.

In figure 1 we show our findings graphically. Figure 1*A* plots classroom average residuals of adjacent sections by professor for introductory and follow-on course achievement as in Kane et al. (2008).[22] Figure 1*B* plots Bayesian shrinkage estimates of the estimated contemporaneous course and follow-on course professor random effects.[23] These results show that introductory course professor value-added in the contemporaneous course is negatively correlated with value-added in follow-on courses (deep learning). On the whole, these results offer an interesting puzzle and, at a minimum, suggest that using contemporaneous student achievement to estimate professor quality may not measure the "true" professor input into the education production function.

C. Results for Follow-on Course Professors

Although the primary focus of our study is to examine how introductory professors affect student achievement, our unique data also allow us to measure how follow-on course professors (e.g., Calculus II professors) affect student achievement in both the contemporaneous course (e.g., Calculus II) and the introductory course (e.g., Calculus I), which should

[21] We cannot directly estimate a standard error for ρ within the random effects framework. Since the denominator, Var $(\lambda_{j^1}^1)$, must be positive, the numerator, ρ Var $(\lambda_{j^1}^1)$, determines the sign of the quotient. As our estimate of ρ Var $(\lambda_{j^1}^1)$ is not significantly different from zero, this result is presumably driven by the magnitude of ρ. Using the two-stage least squares methodology by Jacob et al. (2010) to directly estimate ρ and its standard error, we find it to be negative and statistically insignificant.

[22] Classroom average performance residuals are calculated by taking the mean residual when regressing the normalized score in the course by student on course by semester fixed effects, classroom-level attributes for SAT math, SAT verbal, and academic composite; individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score. In results not shown, we estimate our models using a fixed effect framework as in Kane et al. (2008) and Hoffmann and Oreopoulos (2009) and find qualitatively similar results. To isolate professor value-added from section-specific common shocks in the fixed effect framework, we estimate Var $(\lambda_{j^1}^1)$ and Var $(\lambda_{j^2}^2)$ using pairwise covariances in professor classroom average performance residuals.

[23] The Bayesian shrinkage estimates are a best linear unbiased predictor of each professor's random effect, which take into account the variance (signal to noise) and the number of observations for each professor. Specifically, estimates with a higher variance and a smaller number of observations are shrunk toward zero. See Rabe-Hesketh and Skrondal (2008) for further details.

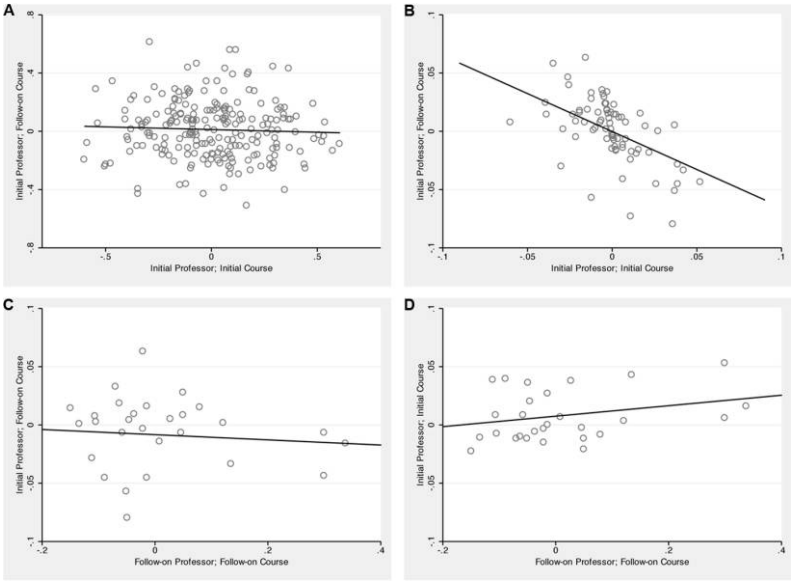

FIG. 1.—Plots of professor effects. *A*, Performance residuals of introductory professor effect in initial course versus introductory professor effect on follow-on course. Classroom average performance residuals were calculated by taking the mean residual when regressing the normalized score in the course by student on course by semester fixed effects, classroom-level attributes for SAT math, SAT verbal, and academic composite; individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score. *B*, Bayesian shrinkage estimates of introductory professor effect in initial course versus introductory professor effect on follow-on course. *C*, Bayesian shrinkage estimates of introductory professor effect of follow-on course versus follow-on professor effect in follow-on course. *D*, Bayesian shrinkage estimates of introductory professor effect in initial course versus follow-on professor effect in follow-on course.

be zero. These results are interesting for two reasons. First, they help test the statistical assumptions of the value-added model as described by Rothstein (2010). Second, we observe a subset of professors in our sample teaching both the introductory and follow-on courses (Calculus I and II). Thus, we are able to examine the correlation between introductory course professor value-added and follow-on course professor value-added.

Rothstein (2010) shows that the assumptions of value-added models are often violated because of the self-selection of students to classrooms and teachers. To illustrate his point, Rothstein (2010) finds that value-added models yield large "effects" of fifth grade teachers on fourth grade test scores. We report estimates for Var $(\lambda^{I}_{j^2})$, the follow-on professor effect on the initial course grade, in row 3, column 3 of table 4. Consistent with random assignment, we find no evidence that follow-on

professors affect introductory course achievement. The estimated variance in the professor random effect is near zero (SD = 0.000002). However, we do find that follow-on professors significantly affect contemporaneous follow-on student achievement. As shown in row 4, column 4, the estimate of $\text{Var}(\lambda^2_{j^2})$ is 0.0185 (SD = 0.136).

To examine the correlation between introductory course professor value-added and follow-on course professor value-added, we show plots of the Bayesian shrinkage estimates in figure 1 for the subset of professors we observe teaching both the introductory and follow-on courses. Figure 1*C* plots fitted values of $\lambda^2_{j^1}$ versus $\lambda^2_{j^2}$ (introductory professor effect on the follow-on course vs. the follow-on professor effect in the follow-on course), and figure 1*D* plots $\lambda^1_{j^1}$ versus $\lambda^2_{j^2}$ (introductory professor effect in the initial course vs. the follow-on professor effect in the follow-on course). These plots yield two interesting findings. First, the clear positive relationship shown in figure 1*D* indicates that professors who are measured as high value-added when teaching the introductory course are also measured as high value-added when teaching the follow-on course. However, the slightly negative and noisy relationship in figure 1*D* indicates that of professors who teach both introductory and follow-on courses, the value-added to the follow-on course produced during the introductory course (deep learning) is uncorrelated with contemporaneously produced value-added in the follow-on course. That is, there appears to be a clear set of professors whose students perform well on sequences of contemporaneous course work, but this higher achievement has little to do with persistent measurable long-term learning.

D. Results for Section-Specific Common Shocks

In both the introductory and follow-on courses, we find significant contemporaneous section-specific common shocks. Although the section-specific common shocks serve primarily to control for section-level variation lest it inappropriately be attributed to professor value-added, the magnitudes and signs of the cross-product common shocks provide a useful check of internal consistency. As expected, the common shock from the introductory course persists into the follow-on course, $\text{Var}(\xi^2_{ts^1}) > 0$. In contrast to Rothstein (2010), the common shock in the follow-on course has no effect on introductory course performance, $\text{Var}(\xi^1_{ts^2}) = 0$. This is further evidence in support of random student assignment into sections with respect to academic ability.

TABLE 5
PROFESSOR OBSERVABLE CHARACTERISTICS AND STUDENT EVALUATIONS OF PROFESSORS

	λ_{j1}^{1} (1)	λ_{j1}^{2} (2)
	A. Professor Observable Attributes	
Associate/full professor	−.69*	.70*
	(.41)	(.40)
Terminal degree	−.28	.38
	(.27)	(.27)
Greater than 3 years' teaching experience	−.79***	.66**
	(.29)	(.29)
	B. Student Evaluation Scores	
Instructor's ability to provide clear, well-organized instruction was	.51***	−.46**
	(.19)	(.20)
Value of questions and problems raised by instructor was	.70***	−.59**
	(.24)	(.25)
Instructor's knowledge of course material was	.56**	−.44*
	(.24)	(.24)
The course as a whole was	.49**	−.39*
	(.23)	(.23)
Amount you learned in the course was	.59**	−.47*
	(.23)	(.24)
The instructor's effectiveness in facilitating my learning in the course was	.54***	−.45**
	(.20)	(.20)

NOTE.—Each row by column represents a separate regression in which the dependent variable is the Bayesian shrinkage estimates of the corresponding professor random effects estimated in eq. (1). In all specifications the Bayesian shrinkage estimates were scaled to have a mean of zero and a variance of one. Panel A shows results for modal rank and mean years of teaching experience. Panel B shows results for sample career averages on student evaluations.

* Significant at the .10 level.
** Significant at the .05 level.
*** Significant at the .01 level.

IV. Observable Professor Characteristics and Student Evaluations of Professors

A. *Observable Professor Characteristics*

One disadvantage of the professor value-added model is that it is unable to measure which observable professor characteristics actually predict student achievement. That is, the model provides little or no information to administrators wishing to improve future hiring practices. To measure whether observable professor characteristics are correlated with professor value-added, we regress normalized Bayesian shrinkage estimates from the contemporaneous course, λ_{j1}^{1}, and follow-on course, λ_{j1}^{2}, on professor observable attributes.[24] Results are presented in table 5, panel A.

[24] For the professor observable attributes we use mean experience and modal rank. We combine the ranks of associate and full professor, as do Hoffmann and Oreopoulos (2009), because of the small numbers of full professors in our sample. Lecturers at USAFA are typically younger military officers (captains and majors) with master's degrees.

The overall pattern of the results shows that students of less experienced and less qualified professors perform significantly better in the contemporaneous course being taught. In contrast, the students of more experienced and more highly qualified introductory professors perform significantly better in the follow-on courses. Here, we have normalized the shrinkage estimates of professor value-added to have a mean of zero and a standard deviation of one. Thus, in column 1, panel A, the negative coefficient for the associate/full professor dummy variable (−0.69) indicates that shrinkage estimates of contemporaneous value-added among professors are, on average, 0.69 standard deviations lower for senior ranking professors than for lecturers. Conversely, the positive and significant result (0.70) for the associate/full professor dummy variable in column 2 indicates that these same professors teach in ways that enhance student performance in follow-on courses. We find a similar pattern of results for the terminal degree and experience variables.

The manner in which student scores are determined at the USAFA as described in Section II.C allows us to rule out the possibility that higher-ranking professors have higher grading standards for equal student performance. Hence, the preceding results are likely driven by the manner in which the course is taught by each professor.[25]

B. Student Evaluations of Professors

Next, we examine the relationship between student evaluations of professors and student academic achievement as in Weinberg, Hashimoto, and Fleisher (2009). This analysis gives us a unique opportunity to compare the relationship between value-added models (currently used to measure primary and secondary teacher quality) and student evaluations (currently used to measure postsecondary teacher quality).

To measure whether student evaluations are correlated with professor value-added, we regress the normalized Bayesian shrinkage estimates from the contemporaneous course, λ^1_j, and follow-on course, λ^2_j, on career averages from various questions on the student evaluations.[26] Results presented in table 5, panel B, show that student evaluation scores are positively correlated with contemporaneous course value-added but negatively correlated with deep learning.[27] In column 1, results for contemporaneous value-added are positive and statistically significant at the

[25] To test for possible attrition bias in our estimates, we examined whether observable teacher characteristics in the introductory courses were correlated with the probability a student drops out after the first year and whether the student ultimately graduates. Results had various signs, were small in magnitude, and were statistically insignificant.

[26] Again, for ease of interpretation we normalized the Bayesian shrinkage estimates to have a mean of zero and a variance of one.

[27] For brevity, we present results for only a subset of questions; however, results were qualitatively similar across all questions on the student evaluation form.

.05 level for scores on all six student evaluation questions. In contrast, results in column 2 for follow-on course value-added show that all six coefficients are negative, with three significant at the .05 level and three significant at the .10 level

Since proposals for teacher merit pay are often based on contemporaneous teacher value-added, we examine rank orders between our professor value-added estimates and student evaluation scores. We compute rank orders of career average student evaluation data for the question, "The instructor's effectiveness in facilitating my learning in the course was," by professor, $r(\omega_j^1) = r_{\omega_1^1}$, and rank orders of the Bayesian shrinkage estimates of introductory professor value-added in the introductory course, $r(\lambda_j^1) = r_{\lambda_1^1}$, and introductory course professor value-added in the follow-on course, $r(\lambda_j^2) = r_{\lambda_1^2}$. Consistent with our previous findings, the correlation between introductory calculus professor value-added in the introductory and follow-on courses is negative, $\text{Cor}(r_{\lambda_1^1}, r_{\lambda_1^2}) = -0.68$. Students appear to reward contemporaneous course value-added, $\text{Cor}(r_{\lambda_1^1}, r_{\omega_1^1}) = 0.36$, but punish deep learning, $\text{Cor}(r_{\lambda_1^2}, r_{\omega_1^1}) = -0.31$. As an illustration, the calculus professor in our sample who ranks dead last in deep learning ranks sixth and seventh best in student evaluations and contemporaneous value-added, respectively.

V. Conclusion

Our findings show that introductory calculus professors significantly affect student achievement in both the contemporaneous course being taught and the follow-on related curriculum. However, these methodologies yield very different conclusions regarding which professors are measured as high quality, depending on the outcome of interest used. We find that less experienced and less qualified professors produce students who perform significantly better in the contemporaneous course being taught, whereas more experienced and highly qualified professors produce students who perform better in the follow-on related curriculum.

Owing to the complexities of the education production function, where both students and faculty engage in optimizing behavior, we can only speculate as to the mechanism by which these effects may operate. Similar to elementary and secondary school teachers, who often have advance knowledge of assessment content in high-stakes testing systems, all professors teaching a given course at USAFA have an advance copy of the exam before it is given. Hence, educators in both settings must choose how much time to allocate to tasks that have great value for raising current scores but may have little value for lasting knowledge.

One potential explanation for our results is that the less experienced professors may adhere more strictly to the regimented curriculum being

tested, whereas the more experienced professors broaden the curriculum and produce students with a deeper understanding of the material. This deeper understanding results in better achievement in the follow-on courses. Another potential mechanism is that students may learn (good or bad) study habits depending on the manner in which their introductory course is taught. For example, introductory professors who "teach to the test" may induce students to exert less study effort in follow-on related courses. This may occur because of a false signal of one's own ability or an erroneous expectation of how follow-on courses will be taught by other professors. A final, more cynical, explanation could also relate to student effort. Students of low-value-added professors in the introductory course may increase effort in follow-on courses to help "erase" their lower than expected grade in the introductory course.

Regardless of how these effects may operate, our results show that student evaluations reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning. Using our various measures of teacher quality to rank-order teachers leads to profoundly different results. Since many U.S. colleges and universities use student evaluations as a measurement of teaching quality for academic promotion and tenure decisions, this finding draws into question the value and accuracy of this practice.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *J. Labor Econ.* 25 (1): 95–135.

Barlevy, Gadi, and Derek Neal. 2009. "Pay for Percentile." Working Paper no. 2009-09, Fed. Reserve Bank Chicago. http://www.chicagofed.org/webpages/publications/working_papers/2009/wp_09.cfm.

Barron, John M., Bradley T. Ewing, and Glen R. Waddell. 2000. "The Effects of High School Participation on Education and Labor Market Outcomes." *Rev. Econ. and Statis.* 82 (3): 409–21.

Bettinger, Eric, and Bridget Terry Long. 2006. "The Increasing Use of Adjunct Instructors at Public Institutions: Are We Hurting Students?" In *What's Happening to Public Higher Education?* edited by Ronald G. Ehrenberg, 51–70. Westport, CT: Praeger.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Educ. Finance and Policy* 1 (2): 176–216.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Estimating Peer Effects in College Achievement." *J. Labor Econ.* 27 (3): 439–64.

Carrell, Scott E., and James E. West. 2008. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." Working Paper no. 14081, NBER, Cambridge, MA. http://www.nber.org/papers/w14081.

Cavalluzzo, Linda C. 2004. "Is National Board Certification an Effective Signal

of Teacher Quality?" Technical Report no. 11204, CNA Corp., Alexandria, VA. http://www.cna.org/documents/CavaluzzoStudy.pdf.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *J. Human Resources* 41 (4): 778–820.

———. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper no. 12828, NBER, Cambridge, MA. http://ideas.repec.org/p/nbr/nberwo/12828.html.

Dee, Thomas S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Rev. Econ. and Statis.* 86 (1): 195–210. http://www.mitpressjournals.org/doi/abs/10.1162/003465304323023750.

———. 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *A.E.R. Papers and Proc.* 95 (2): 158–65.

Ehrenberg, Ronald G., and Dominic J. Brewer. 1994. "Do School and Teacher Characteristics Matter? Evidence from High School and Beyond." *Econ. Educ. Rev.* 13 (1): 1–17. http://ideas.repec.org/a/eee/ecoedu/v13y1994i1p1-17.html.

Ehrenberg, Ronald G., and Liang Zhang. 2005. "Do Tenured and Tenure-Track Faculty Matter?" *J. Human Resources* 40 (3): 647–59.

Ferguson, Ronald F., and Helen F. Ladd. 1996. "How and Why Money Matters: An Analysis of Alabama Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 265–98. Washington, DC: Brookings Inst. Press.

Figlio, David N., and Lawrence W. Kenny. 2007. "Individual Teacher Incentives and Student Performance." *J. Public Econ.* 91 (5–6): 901–14.

Goldhaber, Dan, and Emily Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Rev. Econ. and Statis.* 89 (1):134–50. http://www.mitpressjournals.org/doi/abs/10.1162/rest.89.1.134.

Good, Phillip I. 2006. *Resampling Methods: A Practical Guide to Data Analysis.* 3rd ed. Boston: Birkhauser.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *A.E.R. Papers and Proc.* 61 (2): 280–88.

Harris, Douglas N., and Tim R. Sass. 2007. "Teacher Training, Teacher Quality and Student Achievement." Research Publications, Teacher Quality Research, Tallahassee, FL. http://www.teacherqualityresearch.org/teacher_training.pdf.

Hoffmann, Florian, and Philip Oreopoulos. 2009. "Professor Qualities and Student Achievement." *Rev. Econ. and Statis.* 91 (1): 83–92. http://www.mitpressjournals.org/doi/abs/10.1162/rest.91.1.83.

Jacob, Brian A., and Lars Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." *J. Human Resources* 39 (1): 50–79.

Jacob, Brian A., Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *J. Human Resources*, forthcoming.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Econ. Educ. Rev.* 27 (6): 615–31.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper no. 14607, NBER, Cambridge, MA. http://www.nber.org/papers/w14607.

Koerner, James D. 1963. *The Miseducation of American Teachers.* Boston: Houghton Mifflin.

Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. Texts in Statistics. Secaucus, NJ: Springer.

McCaffrey, Daniel J., J. R. Lockwood, Daniel Koretz, and Laura Hamilton. 2004. *Evaluating Value-Added Models for Teacher Accountability*. Monograph no. 158. Santa Monica, CA: Rand Corp.

Murnane, Richard. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. 2nd ed. College Station, TX: Stata Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, CA: Sage.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *A.E.R. Papers and Proc.* 94 (2): 247–52.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Q.J.E.* 125 (1): 175–214.

Summers, Anita A., and Barbara L. Wolfe. 1977. "Do Schools Make a Difference?" *A.E.R.* 67 (4): 639–52.

Vandevoort, Leslie G., Audrey Amrein-Beardsley, and David Berliner. 2004. "National Board Certified Teachers and Their Students' Achievement." *Educ. Policy Analysis Archives* 12 (46).

Weinberg, Bruce A., Masanori Hashimoto, and Belton M. Fleisher. 2009. "Evaluating Teaching in Higher Education." *J. Econ. Educ.* 40 (3): 227–61.